DOCUMENT RESUME

ED 161 897 95 TM 007 472

AUTHOR Marston, Paul T.; And Others
TITLE Coder Drift: A Reliability Problem for Teacher Observations.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Mar 78
CONTRACT NIE-C-74-0088
GRANT NIE-G-78-0116
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978); For related document, see TM 007479

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Behavior Rating Scales; Classroom Observation Techniques; *Codification; Elementary Secondary Education; *Error Patterns; *Evaluators; Higher Education; *Reliability; Secondary School Teachers
IDENTIFIERS *Coder Drift; Flanders Systems of Interaction Analysis; *Interrater Reliability; Observer Reliability

ABSTRACT

The results of two experiments support the hypothesis of "coder drift" which is defined as change that takes place while trained coders are using a system for a number of classroom observation sessions. The coding system used was a modification of the low-inference Flanders System of Interaction Analysis which calls for assigning observed teacher or student behaviors to ten relatively objective categories. The situations rated in these experiments consisted of videotapes of social studies classes which were conducted by experienced teachers. In each experiment, pairs of coders viewed the same series of ten videotapes, either working with their partners or working in separate rooms. In one condition, the pairs worked together for the first five sessions and apart for the other five. In the other condition, that order was reversed. Pairs working together for the first five sessions showed higher reliabilities than they did when they worked apart, and pairs working apart for the first five sessions showed even lower reliability. The same effect was found for five different methods of computing reliability. Randomly matched individuals showed no drift effects, nor did test-retest reliabilities between pairs working together and those working apart. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED161897

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Coder Drift: A Reliability Problem for Teacher Observations

PAUL T. MARSTON, LINDA K. ZIMMERER, W. CURTIS VAUGHAN

Research and Development Center for Teacher Education

The University of Texas at Austin

TM007 472

This study was supported in part by the National Institute of Education Contract No. NIE-C-74-0088, Grant No. OB-NIE-G-78-0116, Research and Development Center for Teacher Education. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that office should be inferred. Paper presented at the annual meeting of the American Educational Research Association held in Toronto, March 27-31, 1978.

Abstract

Coder drift was defined as the change that takes place in the way an observation system is used after coders have been trained. Two experiments were conducted to test for drift on a low-inference observation system by having pairs of coders view the same videotapes either working with their partner or working in separate rooms. The results confirmed the drift hypothesis and further indicated that the coding of each pair tended to converge on an idiosyncratic norm.

Coder Drift: A Reliability Problem for Teacher Observations

One of the strongest messages that the taxpayers have sent the educators of this country in recent years is the demand for accountability. These citizens are distressed to read newspaper and magazine accounts of high school graduates who cannot read at a level sufficient to hold a job. They see the schools asking for more and more money and yet at the same time appearing to provide less and less education. The answer coming back from the educational community has been an increased emphasis on program evaluation which has as its aim the monitoring of the effectiveness of the various parts of the educational system. These evaluations are intended to identify those parts of the system that are not performing as they should, and either improving or eliminating them.

The classroom teacher represents a key element in the educational system, and, therefore, much of the evaluation effort has been concentrated on his or her performance. A primary source of information about teachers has been the systematic observation of ongoing classroom behavior. This behavior was then related to student performance measures, such as achievement gain, to identify effective teacher behaviors. The results of such process-product studies were then used to make prescriptions about how to teach or how to run a classroom (Bloom, Hastings, & Madaus, 1971; Brophy & Evertson, 1976). While this method clearly represents a straightforward approach to the study of teacher effectiveness, reviewers like Rosenshine (1971) have pointed out that it has not been as successful at identifying effective teacher behaviors as the proponents would like it to be. Part of the problem lies in the methodological shortcomings associated with

using observation as a means of data collection. For instance, there is a major problem with coder reliability, both in terms of how to measure it and in terms of maintaining it at a high level. Rosenshine contends that reliability even means different things to different investigators and that the procedures used to determine interrater reliability may be quite different from study to study. To avoid confusion, interrater reliability in this paper refers to the agreement between two raters who code the same classroom behavior at the same time.

High reliability is essential for observation systems if they are to be an effective research or evaluation tool. This can be achieved at the onset by an intensive training program, but the problem then becomes one of keeping the reliability high throughout the observation period. The usual method is to perform interrater reliability checks at frequent intervals and conduct retraining when necessary. Such checks depend on the key assumption that coders independently deviate from the coding norms established in the training program. It further assumes that the teacher's behavior is relatively constant over successive observations and that a means of classifying all observed behavior was provided in training. Failure to fulfill any one of these assumptions could invalidate the reliability check procedure.

A study by Calkins, Borich, Pascone, Kugle, and Marston (1978) investigated the reliabilities of five low-inference and two high-inference classroom observation instruments in terms of the generalizability theory. According to this theory, any measurement must be considered a sample from a much larger set of possible measurements. In terms of teacher behavior, generalizability depends on the extent to which the codings for teacher

behavior are comparable for coders across other coders and across teaching occasions. Their findings showed that teacher behavior could not be assumed constant across occasions. While Calkins et al. did not study the other two assumptions mentioned above, some of the coder comments they reported call these into question. Their coders made remarks to the effect that they modified the system during the course of their observations. The rationale for making these changes was that they had observed behaviors not covered by the training and had to establish "ground rules" to reliably code them. It appears that each pair of coders may have developed their own, idiosyncratic system for coding classroom behavior. These modifications resulted in more consistent codings between the partners, which means the reliabilities measured between them might be a function of communications occurring during the coding sessions. Such modifications of the system change the coding of specific classroom behaviors away from the normative values.

The coder comments suggested the possibility that the way behavior is coded may change with repeated use of the system. This type of change in the system has been termed "coder drift" for the purposes of this paper. Coder drift can best be studied by having all coders observe the same teacher-classroom situation using a medium such as videotape. A direct approach then would be to compare observations of coders on successive occasions with some normative coding for each tape. Increasing departure from the normative ratings would be evidence for coder drift. Unfortunately, such norms are extremely difficult and time-consuming to construct. A workable alternative is to compare the reliability for coders who communicated with one another about the system with another group of coders who did not talk with one another. This alternative is based on the assumption that coders

who communicate with one another will tend to drift in the same direction because the changes they make are consistent with each other. Coders who do not communicate with one another would tend to drift apart because the changes they each make are independent of each other. A lower reliability figure for coders who do not communicate would then be taken as evidence for coder drift.

To test whether communications with each other would change their observations required the comparison of a situation in which coders talked about the coding with a situation in which they did not talk to one another. The first situation was achieved by having coders work together in pairs, which allowed whatever communications might take place to occur. The second situation was achieved by assigning coders to pairs, but having them work in separate rooms so that they could not communicate with each other during coding.

If this paradigm does provide evidence for coder drift, then it would be interesting to ask if the duration of the drift effect is short- or long-term. A simple test for this is to have the two coding groups switch conditions part way through testing. If the effects are short-term, then the reliabilities should decrease for the pairs switching from together to apart, and increase for the pairs in the group switching from apart to together. If, on the other hand, the drift effect is a long-range one, then the pairs in the group switched from together to apart should not change much, while the pairs in the other group should show an increase.

The Flanders System of Interaction Analysis (Flanders, 1971), which was one of the observation instruments used by Calkins et al. (1978), was selected to study coder drift. It is a widely used, low-inference coding system which was considered to be advantageous for the purposes of

this study because it contained well-defined categories, and a high level of coder reliability could be achieved in a relatively short training period. The ten categories, which describe the interactive behavior of students and teachers, focus on specific, relatively objective behaviors and are recorded at 3-second intervals. The instrument includes seven categories for teacher behavior, two for student behavior, and one additional category for silence or confusion. Flanders originally intended this system to provide feedback to the teacher based on an interpretation of the sequence of teacher-student interactions. The sequencing of behavior did not appear to be useful for the present study, and, therefore, the coding was somewhat simplified.

The coder drift hypothesis was tested in two experiments. The general procedures of both of these were similar, but they differed in the type of classroom setting being observed. In the first experiment the coders observed small, relatively informal classes, while in the second experiment the classes observed were structured in a more traditional manner.

Method

The design of the two experiments was similar. In both experiments the coders saw five videotapes under one of the conditions--working together or working apart--and five more tapes under reversed conditions. The tapes in Experiment 1 were about 50 minutes long and shown in the same order for every coder. In Experiment 2 the tapes were from a different source, were only about 25 minutes long, and were shown in random order for each pair of coders. One additional tape, which had been shown during the training, was included after the fifth session in Experiment 2 for purposes of test-retest reliability.

Coders

The coders for both experiments were paid $4.00 per hour for their training and observation, plus a $20 bonus for being on time for all experimental sessions. The first experiment was conducted during the regular university spring term, and most of the coders came in response to posters displayed in the Department of Education. The second experiment was conducted during the summer term, and most of the coders came in response to a poster in the University Financial Aids Office. Coders in Experiment 1 were mostly undergraduate and graduate education students, while the coders in Experiment 2 came from a much broader range of interests. There were 30 coders to be trained for the first experiment and 28 for the second. One pair in each of the experiments failed to complete the training program, and one pair had to be dropped from Experiment 1 because of habitual lateness on the part of one member. This resulted in 26 coders for each experiment.

Training

The coders were trained as a group to prevent communications between the members of pairs who were to start out coding separately. The training procedures were based on the widely used training system for Flanders developed by Cornell (Note 1). The major modifications consisted of the elimination of specific interpair reliability tests during training and the inclusion of two video training tapes from the same set used in the experimental sessions. The training was carried out in 3- to 4-hour sessions on two consecutive days.

The training procedures for the two experiments were essentially the same. A change was made for the second experiment in order to distribute the training and reliability testing more evenly over the two days. Portions of the training instructions that had confused coders in Experiment 1 were modified for Experiment 2.

The first step in the training was to give the coders a copy of the category descriptions for the Flanders system when they signed up for the experiment. Each coder was asked to memorize the category numbers, titles and descriptions before he or she came to the first training session.

The beginning of the first training session was devoted to procedural matters such as pay vouchers, bonus requirements, and test session scheduling. A pretraining questionnaire was also administered at this time. After the preliminary matters were settled, the Flanders training was started. The coders were first given an outline of the system, and two main uses of the system were stressed: (1) the system could be used to give feedback to the teacher, and (2) it could be used to study patterns of teacher-pupil interactions. The coders were then given a "cover story" for the experiment. They were told that the purpose was to study the problems arising with the Flanders system when there were inconsistencies in coding among persons observing the same teacher behavior and to study the judgments made by coders when these inconsistencies arose. They were told that to make the situation as realistic as possible they would receive "expert" advice on using the system during the two training sessions. After that, they would be "on their own" in much the same way as a person who had been trained and sent out into the field to observe. At this point, the question arose as to what a coder should do if he or she had a question about coding after the experimental phase had started. In Experiment 1 the answer given was "to review the training materials." In the debriefing sessions following the last test tape, some coders expressed the concern that this instruction discouraged communications between partners. To alleviate this misinterpretation, the response given to the same question in Experiment 2 was changed to a suggestion that he or she could "either review

the materials from training or ask your partner about uncertain codings." It was hoped that this new response would lead to the experimental coders acting more like actual ones.

After the cover story had been given, the detailed training began. The first step was to specifically define each of the ten categories, giving specific examples for each. Then a self-test was administered, in which the coders were asked to code statements by teachers and then compare their codings with those done by an expert. Next, the coders listened to an audiotape of a classroom, which was played along with a high-pitched tone every 3 seconds to provide a guideline for proper timing for coding. They coded this tape "mentally," i.e., they only thought about what codes to use as they heard statements, but did not actually write them down. Next, the coders coded three short audiotapes (without the tone signal on the last two). After each tape, written transcripts were used again to compare codings with those of an expert. The last part of the first training session consisted of coding a videotape very similar in length and content to those to be coded in the experiment. A tone was sounded every ten minutes in this tape, indicating the proper time to switch to a new coding sheet.

The second training session consisted of coding three audiotapes and comparing each with the transcripts. Then, a second videotape similar to the test series was shown, and again the tones were inserted every ten minutes. The reliabilities for the coding of this videotape and the one shown at the end of the first training session were used as the basis for forming pairs for the next phase of the experiment.

Videotapes

The videotapes used in the first study were part of a series made for a dissertation by Clark(1976). The teachers and students shown were all volunteers who participated in a study of teacher behavior-student outcome relationships. The teachers--who all had several years experience--taught three 50-60 minute lessons on the same topic from a junior high level social studies unit. A general outline of the presentation order of materials was provided, but each teacher was free to work out his or her own lesson plans. In the study each teacher taught three lessons to each of three groups of eight students who ranged from seventh- to ninth-grade levels. One lesson was selected for each of the teachers for Experiment 1, based on earlier data that showed similar reliability using the tapes. Lessons from each of the two remaining teachers were used for reliability training.

The videotapes for the second experiment were produced for the Individualized Teaching for Effective Coping study (Haak, Lawrence, & Peck, 1971). These tapes were made in public school classrooms at the fifth- and seventh-grade levels for the purpose of providing feedback to teachers. The tapes were made in existing classes, and the teachers were encouraged to be as natural as possible. The tapes were all 25-30 minutes in length. Tapes of ten social studies teachers were selected for the test tapes, and tapes for two other math teachers were selected for use as reliability tapes for training. In Experiment 2 the videotapes were chosen on the basis of quality.

Procedure and Design

In order to test the coder drift effect, two groups of coder pairs with similar reliabilities at the end of training were required. This was accomplished by forming pairs on the basis of reliability during training. Specifically, the reliability was calculated between all possible pairs for the codings of the videotapes at the end of each training session. These reliabilities were transformed using Fisher's r to z method (Winer, 1971), and the mean taken for the two sessions. A trial and error procedure was then used to assign individuals to pairs in order that (1) the pair's mean reliability was close to the average of all possible pairs, and (2) the pair members had sufficient overlap in available times to be scheduled for the testing sessions. After the pairs had been formed, they were randomly assigned to one of two treatment groups. The first group, consisting of six pairs in both experiments, worked together during the first five sessions and apart for the last five sessions. These coders are referred to as the together-apart group. The second group, consisting of seven pairs in both experiments, worked apart during the first part of the experiment and together for the last part. They are referred to as the apart-together group.

An additional question was whether the drift could be observed within an individual coding session. This was tested by dividing each tape into 10-minute segments, which were coded on separate tally sheets. There were five segments in the first experiment and three in the second.

The resultant design had six factors that were: group (2) x coder pair (13) x coder (26) x condition (2) x session (5) x segment (5 or 3).

The experimenters were not the same persons who conducted the training. They were instructed not to answer any questions about how to code the tapes. They played the videotapes on a 19 mm (3/4 in.)

U-matic cassette player. In Experiment 2, two players were used, which allowed different tapes to be shown at the same time. The coders viewed the tapes on a 48 cm (19 in.) black and white television monitor. The coding was recorded on tally sheets which were labeled with the ten Flanders categories across the top and 100 boxes beneath each category. After each 3-second coding interval, a box was checked under the appropriate category. This was not the standard method of coding the Flanders system, but it simplified checking and recording the codes for our purposes. If the number of codes in any category exceeded 100, the coders were instructed to switch to a new sheet. The end of each segment was signalled by an audio tone. In the first experiment these tones were produced by a Sonalert device controlled by a digital timer that was activated by the experimenter when the tape started. On two occasions the timer failed to work and the number of codes had to be estimated as one-half the number recorded because two segments were run together. This problem was corrected in the second experiment by using the timer to record the tone on the videotape sound track. The coding sessions were scheduled every 1 1/2 hours in Experiment 1 and every hour in Experiment 2. It took 17 days to complete the test phase of each experiment.

To determine how much the coders communicated with one another, a microphone was placed above the monitor, and everything the coders said during the sessions was recorded. The instructions led the observers to believe that the microphone was to talk to the experimenter when a button on their table was pushed. This button did turn on a light in the experimenter's room to signal problems with the equipment. The debriefings indicated that most of the coders in Experiment 1 did not know they were being recorded. This was not true in Experiment 2. This difference probably resulted from

a rearrangement of the equipment in the control room part way through the experiment which made it easier to see the recorder from outside the room. In both experiments, the reason for making the recordings, as well as the purpose of the experiment, was explained in the debriefing.

The order of tape presentation was the same for all observers in Experiment 1. Preliminary analysis of the reliabilities indicated there were large differences between the tapes so the tapes in Experiment 2 were shown in random orders. The different orders had the additional benefit of limiting the discussion of coding between members of different pairs. A test-retest reliability session was given to all observers after the fifth session in Experiment 2. This consisted of showing the same tape that was used for reliability coding on the second training day. The coding was done under the same conditions as the first five tapes so that the observers in the together-apart group saw the retest tape while coding together, while those in the apart-together group coded it while working apart. The coders knew in advance that they would be seeing this tape again.

Results

The raw data obtained from both experiments were the frequencies of occurrence of each of the ten Flanders categories during each coding segment. No analyses of the raw frequencies are reported because it is not clear how they should change if drift occurred. The focus of the test of the coder drift hypothesis was on the agreement within a pair of coders, so two derived measures of observer agreement were used for dependent variables: the interrater reliability between pair members,and the absolute difference in codings between pair members.

A reliability measure gives a single value for the agreement of one pair of coders during one segment of one tape. Its value can range from zero, indicating chance agreement between the pair members, to one, indicating perfect agreement. There are several methods of calculating reliability, so--rather than risk having the results be dependent on the measure used--all analyses were repeated with five different measures of reliability. Two of the measures were based on the proportion agreement between the coders, and the other three were correlational. The proportion measures were taken from Scott (1955), and were simple proportion agreement (PO) and proportion agreement correction for chance (PI). The correlation measures were taken from Winer (1971), and include the Pearson product moment correlation (R), the intraclass correlation (RIM) and the intraclass correlation with anchor points (RIA).

One concern about doing an analysis on measures of reliability is the non-normality of their expected distributions. In case this was a problem, the analyses were repeated using transformed scores. The arcsin transformation was used for the proportional measures (PO and PI) and the Fisher r to z transformation was used with the correlational measures (R, RIM, and RIA) (Winer, 1971). In all cases, the findings were of the same form for all measures of raw and transformed reliability. Only the results for Scott's coefficient (PI) and intraclass correlation (RIM) are reported to simplify the discussion.

The second measure of agreement was the absolute difference in frequency of category usage between pair members during a segment. This allows the reliability contribution of individual categories to be measured. This measure is sensitive to the absolute rate of coding, but adjusting it for coding rate causes other problems. The main one is that coders who work at different rates can never "agree" on low-frequency categories.

All of the analyses used the same mixed model analysis of variance that had a between-pair factor coding group and within-pair factors condition, session, and segment. The group and condition factors were considered fixed, while the other three--pair, session, and segment--were considered random. The pair factor was considered nested in the group factor, the session factor was nested in the condition factor, and the segment factor was nested in both the condition and session factors. The complex, random nature of the within-pair variables was necessary because each segment of classroom behavior was unique. Each teacher had his or her own way of presenting material, even when the topic was the same across teachers. The random variables in this design make direct F-tests impossible for many effects, including the ones of interest. So, a quasi F-ratio, F", described by Winer (1971), was used to test those effects.

There are two effects of interest. The main effect is for group, which tests reliability differences between pairs that start out coding together and switch to apart, and those that start out coding apart and switch to together. If this effect is significant, then a relatively persistent change has occurred as a result of the initial coding conditions. The other effect of interest is the group x condition interaction. If it is significant, then there was an immediate change in coder performance as a function of coding conditions. The complex nature of the design makes it difficult to interpret the session or segment effects, so these are not reported. Suffice it to say there were no clear trends in either variable in either experiment.

All of the reliability measures, both transformed and untransformed, showed a significant ($p < .05$) group x condition interaction in each of the experiments. Each group showed a higher mean reliability when coding

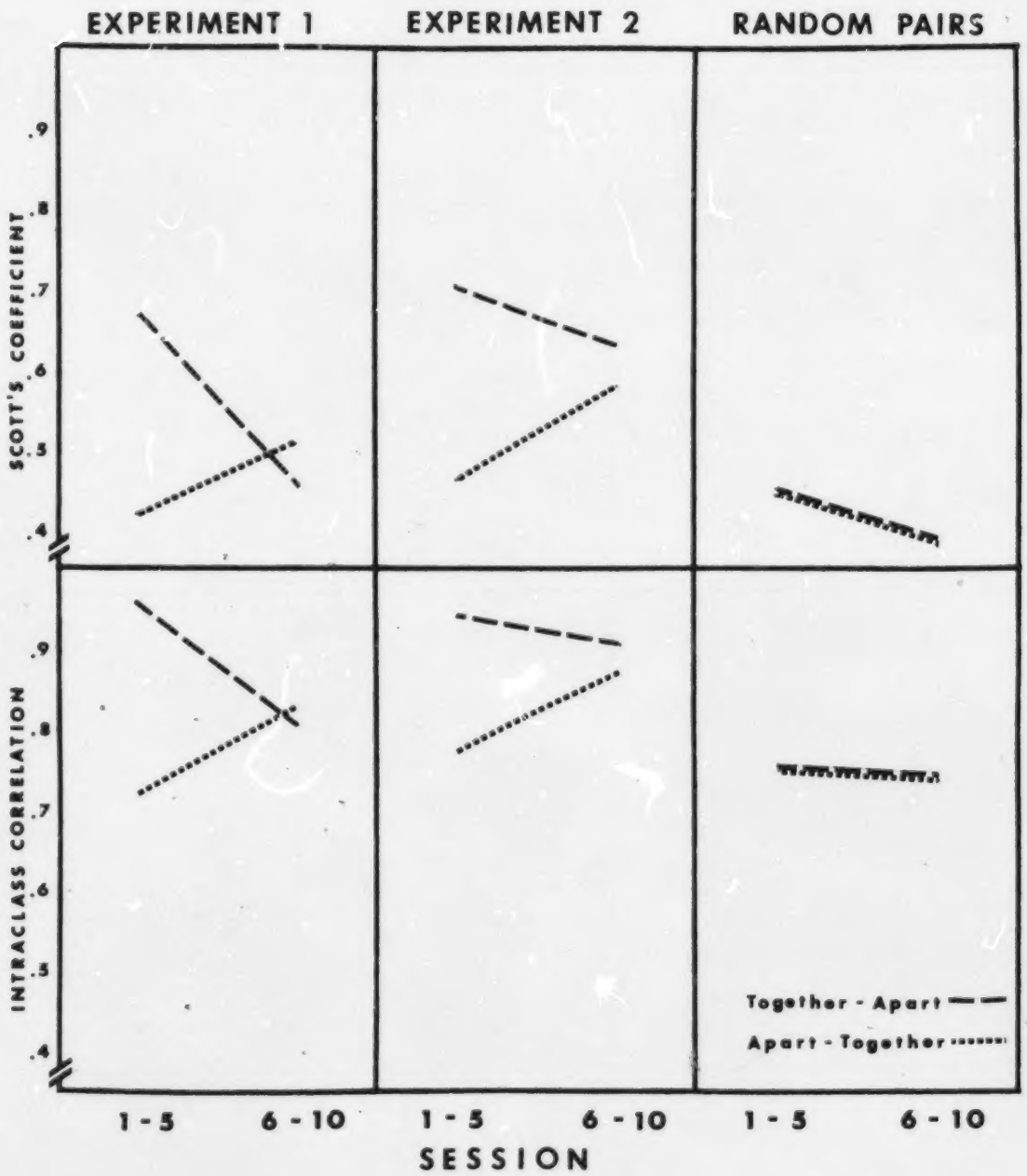

Figure 1. Mean reliability measures for observer pairs as a function of whether the pairs in the group worked together or separately and in what order this occurred.

together than when coding apart. Figure 1 shows this interaction for the PI and RIM measures. In Experiment 1, the interaction test for PI was highly significant, F'' (1, 17) = 10.54, $p < .01$, and so was the interaction for RIM, F'' (1, 19) = 8.19, $p < .01$. In Experiment 2, both the interaction test for PI, F'' (1, 19) = 5.16, $p < .05$, and the test for RIM, F'' (1, 16) = 6.27, $p < .05$, were significant. A main effect for groups was also observed in all measures for Experiment 2, but it failed to reach significance for two untransformed reliability measures (PI and R) and one transformed measure (PI). In all three cases the F''-ratio was close to significance ($p < .10$). The significant main effect tests for group in Experiment 1 were PI, F'' (1, 19) = 3.56, $p < .10$, and RIM, F'' (1, 19) = 9.73, $p < .01$. For Experiment 2 the corresponding tests were RI, F'' (1, 15) = 5.56, $p < .05$, and RIM, F'' (1, 14) = 7.01, $p < .05$. The means and F''-ratios for all ten measures are shown in Table 1.

To test whether the mean reliability increases shown for pairs working together were the result of more accurate coding, random pairs were formed within groups and then given the same analysis as the actual pairs. Only the data from Experiment 1 could be used for this test because the order of presentation differences in Experiment 2 precluded forming the pairs. Neither the group effect nor the interaction of group x condition was significant for any of the measures using random pairs. In fact, the largest F'' found was .69. The right hand figures in Table 1 show what little difference there was between the groups made of random pairs.

The analysis of variance of the absolute differences in category frequency showed that most of the reliability effect was the result of two categories: "giving directions" (6) and "student talk--response" (8). In Experiment 1 the group x condition interaction for category 6

Table 1

Reliability Analysis for Experiments 1 and 2

Experiment	Measure[1]	Group[2]					Group x Condition						
				F"			T-A		A-T		F"		
		T-A	A-T	df_g		Value[3]	Together	Apart	Apart	Together	df		Value
1	PO	.676	.586	1	18	4.45*	.767	.584	.544	.628	1	17	10.76**
	PI	.559	.462	1	19	3.76	.664	.455	.422	.502	1	17	10.54**
	R	.888	.820	1	18	3.72	.961	.814	.787	.853	1	16	10.73**
	RIM	.877	.775	1	18	6.37*	.953	.801	.727	.823	1	19	8.19**
	RIA	.874	.779	1	17	5.75*	.952	.796	.738	.820	1	18	8.69**
2	PO	.753	.623	1	15	6.42*	.783	.724	.575	.672	1	17	5.33*
	PI	.669	.521	1	15	5.46*	.708	.631	.465	.576	1	19	5.16*
	R	.945	.841	1	13	7.05*	.961	.930	.799	.884	1	15	4.76*
	RIM	.922	.821	1	14	7.15*	.942	.902	.770	.871	1	16	6.27*
	RIA	.925	.818	1	13	7.36*	.945	.904	.769	.868	1	16	6.20*
1 Random Pairs	PO	.556	.544	18	16	.23	.581	.532	.566	.521	89	17	.45
	PI	.415	.411	87	16	.15	.443	.387	.441	.381	89	18	.38
	R	.815	.787	3	16	.48	.848	.781	.807	.767	18	18	.69
	RIM	.745	.744	88	16	.22	.757	.733	.752	.737	86	19	.48
	RIA	.752	.745	67	16	.25	.770	.735	.760	.731	89	18	.53

Notes: [1]Reliability measure - see text for definition of abbreviation.

[2]T-A - Together-Apart Group
A-T - Apart-Together Group

[3] *$p < .05$
**$p < .01$

was significant, F'' (1, 19) = 6.13, p <.05, as was the same test in Experiment 2, F'' (1, 19) = 8.25, p <.01. For category 8 this same interaction was significant in Experiment 1, F'' (1, 14) = 5.11, p < .05, and in Experiment 2, F'' (1, 19) = 4.67, p <.044. In all cases, the absolute difference in number of errors was less for the groups working together. Only one significant main effect for groups was observed on the absolute difference measure, and that was for category 9 in Experiment 2, F'' (1, 18) = 6.38, p <.05. This effect did not occur in Experiment 1, F'' (2, 19) = 1.21, p <.10. A summary of the absolute difference findings is shown in Table 2. When the absolute difference analysis was run on the random pairs, no significant F''-ratios were found. The magnitude of the differences in the random pairs was about the same as those corresponding differences in the actual pairs working apart.

The only significant findings for the test-retest reliability given in the experiment involved the segment effect for the PI measure, F'' (2, 48) = 10.83, p <.01. The coders showed a drop in reliability from the first to the last segment, with means of .753, .709, and .617 for segments 1, 2, and 3 respectively. When compared with the mean reliability of .642 for the together groups, the coders did not do much better with themselves than with another coder. The higher reliability in the first segments may have resulted from memory of the training tape.

Discussion and Conclusions

The results of the two experiments support the hypothesis of coder drift. It was shown both within individual pairs, which showed higher reliability when coding together instead of apart, and in comparison with other pairs which showed the higher reliability for those pairs working together. The drift effect was robust in that it did not depend on the

Table 2

Absolute Difference Analysis for Experiments 1 and 2

Experiment	Category[1] #	Category[1] Mean Frequency	Group[2] T-A	Group[2] A-T	Group[2] F" dfs		Group[2] F" Value[3]	Group x Condition T-A Together	Group x Condition T-A Apart	Group x Condition A-T Apart	Group x Condition A-T Together	Group x Condition F" dfs		Group x Condition F" Value
1	1	.1780	.163	.300	2	13	2.45	.293	.0333	.474	.126	17	14	1.28
	2	1.769	1.25	1.74	1	16	1.61	1.11	1.39	1.90	1.58	2	17	.98
	3	13.46	7.84	9.24	2	14	.61	8.51	7.17	11.5	6.95	1	18	1.96
	4	25.68	5.81	8.77	1	13	.96	5.91	5.71	12.6	4.91	1	15	3.03
	5	57.39	12.0	18.7	1	19	3.22	10.1	13.9	18.1	19.3	33	18	.50
	6	10.31	5.72	5.43	19	18	.22	3.43	8.01	5.88	4.99	1	19	6.13*
	7	.2696	.343	.283	6	19	.51	.167	.520	.200	.366	3	15	1.21
	8	43.72	15.6	26.1	1	19	2.88	9.66	21.5	31.7	20.4	1	14	5.11*
	9	46.25	19.7	25.7	2	19	1.21	9.95	29.4	36.5	14.8	1	17	15.17**
	10	16.12	8.25	10.6	2	18	.63	6.16	10.3	13.3	7.92	1	18	2.15
2	1	.05992	.0833	.0762	82	15	.93	.0444	.122	.0857	.0667	2	19	1.42
	2	1.637	.950	1.79	1	12	.90	.856	1.04	1.25	2.32	2	15	.86
	3	21.81	8.51	13.4	1	13	2.72	6.83	10.2	11.9	15.0	89	16	.40
	4	34.84	7.31	11.5	1	18	3.39	7.06	7.56	12.3	10.8	5	11	1.20
	5	47.43	11.1	16.1	1	17	4.64*	10.0	12.2	19.5	12.7	1	16	3.54
	6	19.39	8.43	7.39	2	16	.86	7.80	9.06	9.19	5.58	1	19	8.25**
	7	.2256	.228	.290	8	18	.67	.311	.144	.305	.276	7	16	.85
	8	60.43	14.0	17.5	2	19	.75	10.5	18.6	21.7	13.3	1	19	4.67*
	9	7.989	4.01	10.6	1	18	6.38*	3.22	4.79	12.2	8.98	3	17	1.19
	10	12.71	4.52	5.87	2	19	1.11	4.10	4.94	7.92	3.82	1	19	4.30*

[1] 1-Accepts Feeling; 2-Praises or Encourages; 3-Accepts or Uses Ideas of Students; 4-Asks Questions; 5-Lecturing; 6-Giving Directions; 7-Criticizing or Justifying Authority; 8-Student Talk--Response; 9-Student Talk--Initiation; 10-Silence or Confusion.

[2] T-A - Together-Apart
A-T - Apart-Together

[3] *$p < .05$
**$p < .01$

type of reliability measure or the use of a data transformation.

As previously discussed, there are two ways in which a pair of coders could maintain a higher interrater reliability. One way would be for them to correct each other so that they stayed as closely as possible to the norm established in training, while the second way would be to shift their coding toward a common norm. In the former case, all coders working together would tend to stay reliable with each other, while in the latter case, only the partners would be reliable. The second, idiosyncratic method of achieving high reliability would not be as desirable because it does not maintain the accuracy of the observation system. If it does occur, coders trained at different times or places would use a system in a different manner, and this, in turn, would limit the chances of replicating results across studies.

The results of three analyses in this study favored the idiosyncratic interpretation of coder drift. The strongest support came from the analysis of random repairings within groups in Experiment 1. These pseudo-pairs showed no drift effects and had average reliability levels close to the real pairs when working apart. The second piece of evidence came from the analysis of the test-retest reliabilities in Experiment 2. Again, there was no reliability difference between the pairs coding apart or together, and the test-retest reliability was closer to interpair reliability of coders working apart. The third source of evidence was the loss of reliability for coder pairs when they were switched from working together to working apart. The first two findings provide strong support for the idiosyncratic interpretation of coder drift. While the third finding is not as strong a test, its correspondence adds weight to the hypothesis.

The results indicated coder drift had both long-term and short-term components. When pairs were switched from coding apart to together, they

increased their reliability, but it never reached the level of reliability for the pairs initially coding together. Conversely, when pairs who initially worked together were switched to coding apart, they never dropped to the reliability level of the pairs who started out coding apart. Because the initial coding condition does effect overall performance, it must be assumed that part of the effect of working together was a long-term effect on the way the coders used the observation system. The short-term components of coder drift were indicated by the rapid change in reliability when the coders switched from one condition to another. Most of this change occurred in the first session and was always completed by the second one. The relative magnitude of the short-term to long-term drift effect varied between experiments. The short-term drift was greater in Experiment 1, as indicated by the higher reliability of the apart-together group than of the together-apart group on the second set of tapes. This did not happen in Experiment 2, where the reliabilities of the together-apart group were always higher. There was, however, a greater gain in reliability for the apart-together group.

It is interesting to speculate on the possible underlying causes of coder drift. One possibility for the long-term changes in coding may be the result of reviewing the training materials for coding after each session, or another possibility could be establishing new ground rules for coding situations that have not been seen before. The former process would tend to keep coding close to the system norms, while the latter might lead to idiosyncratic norms for each pair. Perhaps the short-term effects on reliability could be due to verbal and nonverbal communications between the coders during observation. This would include noticing the other person's statements on how a behavior should be coded, his or her reaction

to a behavior being observed, and noticing his or her rate of coding. Evidence for communication effects is currently being collected from the audiotapes and coder interviews and will be presented as a separate report (Kugle, 1978).

The communication effects described above are related to the larger process of group influence on the formations of norms and attitudes. Sherif (1958) studied group influence on the perceived size of the autokinetic effect. This effect involves the perceived movement of a stationary point of light shown in an otherwise totally dark room. Since this task is ambiguous, the amount of perceived motion varies between individuals, but tends to be constant for each person. Sherif found that the range and norm for the effect, formed by individuals working alone, tended to converge on a common value when they were switched to group observations. This convergence was not as great as that found for individuals who had not had the opportunity to establish a range and norm outside of a group. The results of the current study are strikingly similar, even though here a group consists of only a pair of coders. Those coders who start out working together showed the greatest convergence, as evidenced by their reliability, while the convergence was not as great for those who had a chance to develop ranges and norms by coding apart.

More recent work on conformity has shown that group pressure can change judgments, even when the task is not ambiguous (Asch, 1956; Deutsch & Gerard, 1955). These experiments involved having group members make judgments about the length of a line. Only one member of the group was an actual subject; the others were confederates of the experimenter who gave obviously wrong answers. Of the factors found to be important for conformity,

three were present in the drift study. First, the amount of conformity was greater when the group worked in a face-to-face situation, which corresponds to the coders working together. Second, the comformity was reduced when the subject had to make a written commitment prior to hearing the decisions of the others, which corresponds to coders recording their immediate judgments. Third, the amount of conformity was greater when the subject had to rely on less certain information, which corresponds to coders being faced with new teacher behaviors. Deutsch and Gerard concluded that normative social influences could exert strong pressures to conform, unless the group encouraged the individuals to express their own judgments. The pressure to achieve high reliability with a partner found in most training situations could be a source of conformity.

Close examination of the Flanders categories shows the greatest changes in reliability cannot be fully explained by the notion of conformity. In the case of "teacher giving directions," category (6), the conformity may be the result of a difficult judgment. The coders did not make a clear separation between this behavior and the much more frequent "teacher lecturing," category (5). This type of explanation does not hold up for "student talk--response," category (8), which also indicated drift, but occurred more often than the other student category, "student talk--initiated," category (9). The explanation of drift because of ambiguity may still hold because the coders reported a great deal of difficulty in coding student talk from videotapes. In Experiment 1, where there was more overall student talk, and, in particular, much more student-initiated talk, both categories 8 and 9 showed coder drift. It appears that further experiments will be necessary to isolate the mechanism for the coder drift effect.

The Flanders system presented a relatively strong test of coder drift. Compared to other rating systems, it presents a set of well-defined categories that are easy to code (Calkins et al., 1978). The constant and frequent coding procedure forces coders to concentrate on their own judgments. Despite these factors operating to achieve independent and accurate codings, there was evidence of communication within observer pairs.

Three conclusions on how to improve the accuracy of coders' performance can be drawn from these experiments. First, it is essential that coders be trained to a normative standard on how the system is to be used. The practice of using interrater reliability for a group of coders is insufficient to show that they agree with other users of the system; it only shows that they have conformed to a common norm for coding that system. A second implication is that observer reliability should be tested against some normative standard once the system is in use, and retraining should be undertaken when the reliability with this standard drops below some predetermined value. Third, the differences in findings across the two experiments suggest that one coding system may be more suitable than another for one type of classroom situation. The range of effective situations for each observation system needs to be identified in future research. By carrying out these three implications, researchers should be able to obtain more accurate observations, which should then lead to a greater comparability across studies using the same observation system.

Reference Note

Cornell, T. EPIC Systems, 4901 East Fifth Street, Tuscon, Arizona 85732.

References

Asch, S.E. Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological Monographs, 1956, 70, (9, Whole No. 416).

Bloom, B.S., Hastings, J.T., & Madaus, G.F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971.

Brophy, J.E., & Evertson, C.M. Learning from Teaching: A Developmental Perspective. Boston: Allyn and Bacon, 1976.

Calkins, D., Borich, G.D., Pascone, M., Kugle, C.L., & Marston, P.T. Generalizability of teacher behaviors across classroom observation systems. Journal of Classroom Interaction, 1978, 13, 9-22.

Clark, C.M. The effects of teacher practice on student learning and attitudes in small group instructions. Technical Report No. 47. Stanford Center for Research and Development in Teaching, February, 1976.

Deutsch, M., & Gerard, H.B. A study of normative and informational social influences upon individual judgment. Journal of Abnormal and Social Psychology, 1955, 51, 629-636.

Flanders, N.A. Analyzing Teaching Behavior. Reading, Mass.: Addison-Wesley, 1971.

Haak, A., Lawrence, W.A., & Peck, R.F. ITEC: An experiment in school consultation. Publication No. 1082, Research and Development Center for Teacher Education, The University of Texas at Austin, 1971.

Kugle, C.L. A potential source of bias in classroom observation systems: Coder drift. Paper presented at American Educational Research Association, Toronto, Canada, 1978.

Rosenshine, B. Teaching Behaviours and Student Achievement. Windsor, Berks, Great Britain: National Foundation for Educational Research in England and Wales, 1971.

Scott, W.A. Reliability of content analysis: the case of nominal scale coding. The Public Opinion Quarterly, 1955, 19, 321-325.

Sherif, M. Group influences upon the formation of norms and attitudes. Maccoby, E.E., Newcomb, T.M., & Hartley, E.L. (Eds.) Readings in Social Psychology (3rd Ed.) New York: Holt, 1958.

Winer, B.J. Statistical Principals in Experimental Design (2nd Ed.). New York: McGraw-Hill, 1971.